Wake-Sleep Energy Based Models for Continual Learning

Vaibhav Singh
Concordia University, Mila
vaibhav.singh@mila.quebec

Anna Choromanska
New York University
ac5455@nyu.edu

Shuang Li
University of Toronto
shuang.li@utoronto.ca

Yilun Du
Massachusetts Institute of Technology
yilundu@mit.edu

Abstract

This paper introduces a novel approach for continually training Energy-Based Models (EBMs) on the classification problems in the challenging setting of class incremental learning. Despite the fact that EBMs offer longer retention of knowledge on prior tasks, training EBMs contrastively remains a challenge. Driven by biological plausibility, we leverage the observation that sleep in humans supports active system consolidation and propose a new approach for training EBMs, which we call Wake-Sleep Energy Based Models (WS-EBMs), which rely on wake-sleep cycles. Our training approach consists of short wake phases followed by long sleep phases. During the short wake phase, the free energy associated with ground truth labels is minimized, which conditions the model towards the correct solutions. This is followed by a long sleep phase, where the free energy of the whole system is minimized contrastively, which allows the model to push the energy of incorrect solutions further from the correct response. We provide a theoretical analysis of WS-EBM showing that it satisfies the sufficient condition for designing proper EBM loss. Our empirical evaluation confirms the plausibility of our approach and demonstrates favorable performance of WS-EBM compared to traditional EBM training as well as state-of-the-art class-incremental continual learning techniques. Furthermore, our proposed two-phase training strategy can be easily integrated with existing techniques resulting in substantial boosts in their performance. Finally, we also provide interesting insights justifying our approach by analyzing the orthogonality between the sequential task vectors, and flatness of the optimized energy surfaces, which may guide the design of class incremental continual learning strategies.

1. Introduction

Modern Deep Learning algorithms can be viewed as isolated single-task learning methods trained on data samples that are assumed to be independent and identically distributed (i.i.d). Single-task learning schemes are not equipped with mechanisms allowing them to transfer knowledge, or incrementally learn under data distributional changes or when a new task comes. Therefore they suffer from a phenomenon known as catastrophic forgetting [34]. Humans on the other hand learn continually and accumulate knowledge over time through sense perception [49]. A variety of approaches have been proposed to mitigate catastrophic forgetting, like using regularization-based methods [25, 33, 46], external memory [30, 32], and dynamic model architecture techniques [47].

This paper explores the use of Energy-Based Models (EBMs) for continual learning and proposes an efficient training regime that helps alleviate the problem of catastrophic forgetting. EBMs offer considerable freedom to choose what classes to update in the continual learning process. They look at classification problem from the lens of training an un-normalized probability distribution, which leads to significant improvements in the performance on the classification problems in the continual learning setting [29]. Our work focuses on the setting, where the model architecture is fixed, as opposed to the dynamic architecture techniques or methods incorporating attention or fusion modules [12, 31, 55, 56, 58]. Recent developments [13, 18, 38] in training large-scale EBMs parameterized by deep neural networks on high-dimensional data has motivated us to explore them in the scenario of continual learning framework. In lieu of classification problems, given the input x and output class y, the objective in training an EBM is to shape an energy function $E(w, x, y)$, parameterized by weights w, in such a way that the model produces the correct class label y from a set of possible classes $\mathcal{Y}$ when the energy function E attains its minimum [28]. To train EBMs

we need to define a loss functional $\mathcal{L}(E, x, y)$, which determines the quality of the Energy Function E. The learning objective becomes minimization of this loss functional as $w^* = \min_{w \in \mathcal{W}} \mathcal{L}(E, x, y)$.

EBMs offer a unique perspective into addressing the continual learning problem. With EBMs, continual learning problems simply correspond to constructing an energy landscape that assigns low energy to correct classes and high energy to incorrect classes. EBMs give us freedom in framing the continual learning problem, where we may decide to learn about new classes, by decreasing the energy of that class or choosing to forget prior classes, by increasing the energy of incorrect classes.

In this paper, we continually learn with EBMs by using two separate wake-sleep phases in training. This approach is biologically plausible since in the human brain, deep sleep supports active system consolidation [6]. Although vast amounts of information activate the brain during a daytime period of wakefulness, aggregation and long-term encoding of this information happen during sleep. A global strengthening of newly acquired memory traces and underlying synaptic connections during any single-phase consolidation would inevitably result in a system overflow [6, 57]. Therefore single-phase learning is neither biologically plausible nor efficient, as opposed to relying instead on decoupled phases for knowledge aggregation. Two-phase learning indeed seems to be a crucial adaptive function of active memory consolidation in biological systems. Our proposed two-phase approach consists of, what we call, a wake phase and a sleep phase. In the short wake phase, we minimize the free energy on real data without any contrastive sampling. This is followed by long wake cycles where learning happens through the minimization of the free energy of the whole system by optimizing a suitable loss function. We hypothesize that the model parameters, as shown in Figure 1, during the wake phase capture the ground truth labels by freely minimizing their energies without any constraints. This gives the model conditioning over the correct class labels. But simply running the wake phase for the entire training procedure will deprive the model of learning about the incorrect class labels, leading to a mode collapse [14, 28]. This is prevented by introducing a longer sleep phase, which provides a margin for the model to push up the energy of the incorrect solution from the already conditioned model obtained in the wake phase.

The contributions of this paper can be summarized as:

- We propose a novel and simple framework called ***Wake Sleep Energy-Based Model (WS-EBM)***, which offers an effective way of training EBMs motivated by biological plausibility.
- We apply our training approach in the challenging scenario of class incremental continual learning.
- We offer a theoretical understanding supporting our algorithm.
- We demonstrate that our proposed method outperforms other established baselines and provide some interesting insights justifying our approach.

2. Related Work

One of the key challenges in continual learning is to mitigate the problem of catastrophic forgetting [34]. The primary objective of the model is to adapt to changes in the distribution of the input data while retaining the previously learned knowledge or at least demonstrating graceful degradation. [51] outlines different settings where the model at hand is supposed to solve the classification problem when successively learning tasks. Firstly, we have task incremental setting (Task-IL or multi-head) [16], where the model is trained on data coming from the current task and has access to the task identity of test samples at the inference time. Next, we have domain incremental setting (Domain-IL), where the model incrementally learns a set of tasks, but with the crucial difference that at least at the test time, the trained model does not have any information about the identity of the task that a currently observed sample belongs to. Moreover, identifying the task is not necessary, because each task has the same possible outputs (i.e., the same classes are used in each task), and the changes occur only in the input distribution [35]. Finally, we have class incremental scenario (Class-IL or single-head) where the model requires the task identifier to be predicted along with the class label [41, 47, 54].

Task-IL is the easiest setting to address. Methods dedicated to tackling Task-IL typically employ multi-headed architecture [45]. For Domain-IL algorithms use a single head architecture to classify the input [35]. Lastly, in Class-IL, existing methods need to store data, use replay, or pre-train models on another large data set [4, 20, 33, 42, 44] to perform well in this setting. Recent works in continual learning have focussed more on the general and most challenging scenario of class incremental setting [1, 4, 22, 37]. Existing approaches can be divided into two groups: regularization-based and rehearsal-based methods. In the former, regularization-based terms are specifically used to maintain a balance between stability and plasticity. Typically a penalty is introduced that prevents modifications in the weights of the model that are crucial for the previous tasks while learning the current task at hand. Often these methods are effective for short sequences of tasks but are hard to scale to more difficult problems [1, 16]. The rehearsal-based methods leverage a memory buffer to store examples from previous distributions. Experience Replay (ER) [17, 43] simply replays the stored examples along with the input stream to simulate training over an independent and identically distributed task (joint training). Despite its simplicity, this the method has proven to be highly effective even with a minimal memory footprint [9] and serves as

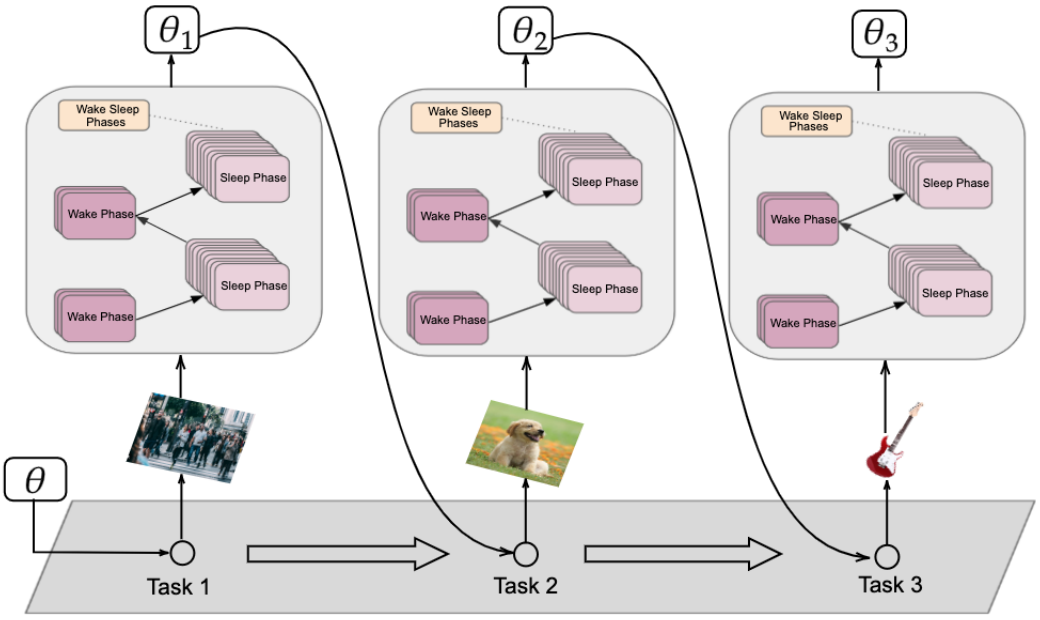


Figure 1. Overview of our Wake-Sleep algorithm for training Energy-Based Models(WS-EBM) for continual learning over image classification in the class incremental setting. Model Parameters(θ) are trained through the alternate short wake and long sleep cycles and are fed to the next task.

the basis for recent methods that propose modifications to the strategy for selecting samples that should be included in the memory buffer [3] or design strategies for sampling the examples from the memory buffer [9, 48]. Finally, the retained knowledge can also be used as a mean to revise the optimization procedure: MER [23] employs meta-learning to discourage interference and maximize knowledge transfer between tasks, while GEM [32] and A-GEM [8] use old training data to minimize the gradient interference in an explicit fashion.

3. EBMs for classification problems

In this work, we focus on the class incremental setting [4, 20, 33, 51], where the model at inference chooses between the classes from all tasks seen so far to predict the label of an input data. In a class incremental setting, the model is trained for a classification task from a stream or sequence of data partitioned into distinct sections where each section holds different non-overlapping class groups. In the t_{th} section, the classifier is fed the training dataset $D_t = (x_i^t, y_i^t)_{t=1}$, where x_i^t , y_i^t are input samples and the corresponding labels, respectively (subscript i denotes the index of the data point in the t_{th} task). This D_t is not accessible later when training on the other data sections. Upon training on the t_{th} section, the model is evaluated across all the class labels seen till now. In other words, training happens only on a single section of data but the model is tested for all the classes that the model has been trained on. Further in this work we assume that task boundaries are known [25, 47, 59] at training time but are not available at inference. Due to the simplicity of EBMs, they can be easily extended to the boundary-free setting [1, 42, 60]. When solving the classification problem via EBMs, the conditional likelihood of a label y given x is sampled from a Boltzmann distribution

$$p_w(y|x) = \frac{exp(-E(w,x,y))}{Z(w;x)}$$
$$\text{where} \quad Z(w;x) = \sum_{\bar{y}\in\mathcal{Y}} exp(-E(w,x,\bar{y})) \tag{1}$$

$\mathcal{Y}$ is discrete set of possible class labels and $E(w,x,y) : (\mathbb{R}^D, \mathbb{N}) \rightarrow \mathbb{R}$ is the energy function that maps an input-label pair (x,y) to a scalar energy value. $Z(w;x)$ is the partition function used for normalizing the distribution. It is desired that the distribution defined by the energy function $E(w,x,y)$ captures the data distribution p_D. This can be done by minimizing the negative log-likelihood of the data, $\mathcal{L}(w)$ defined as follows:

$$\mathcal{L}(w) = \mathbb{E}_{(x,y)\approx p_D}[-log(p_w(y|x))] \tag{2}$$

Expanding the probability distribution we get

$$\mathcal{L}(w) = \mathbb{E}_{(x,y)\approx p_D}[E(w,x,y) + log(\sum_{\bar{y}\in\mathcal{Y}} e^{-E(w,x,\bar{y})}] \tag{3}$$

Directly maximizing the free energy over all labels restricts the model by penalizing all the classes equally. To alleviate this, Equation 3 can be approximated via contrastive

divergence loss [13, 21]:

$$\mathcal{L}_{CD}(w) = \mathbb{E}_{(x,y)\approx p_D}[E(w,x,y) - E(w,x,\bar{y})], \quad (4)$$

where y is the ground truth label of data x and $\bar{y}$ is a negative class label randomly sampled from the set of class labels in the current training batch $\mathcal{Y}_b$ such that $\bar{y} \neq y$.

4. Wake-Sleep Energy Based Model(WS-EBM): Proposed Strategy

The contrastive divergence loss function in Equation 4 requires two terms. $E(w, x_i, y_i)$ is the energy of the ground-truth label, called the positive energy. And $E(w, x_i, \bar{y_i})$ is the negative energy corresponding to the mismatch between the input x_i and the label $\bar{y_i}$ that is defined as the most offending answer with the lowest energy among all the incorrect labels. Since each task in our experiments is a two-class classification, there is only 1 most offending label for each data point x_i.

Algorithm 1 Wake sleep training of EBMs in class incremental setting

Require: Data $D = (x_i, y_i, T_t)$, *Iterations*(iters), *Tasks*(tasks), T_t is the current task id, $\mathcal{L}$ is the contrastive divergence loss. *Wake Cycles*(wc), *Sleep Cycles*(sc)

for t in tasks **do**
  for i in iters **do**
    for w_c in wc **do**
      $L_{wc} = E(w, x_i, y_i)^2$
      $w = w - \eta\nabla(L_{wc})$
    end for
    for s_c in sc **do**
      $L_{sc} = \mathcal{L}(E(w, x_i, y_i), E(w, x_i, \bar{y_i}))$
      $w = w - \eta\nabla(L_{sc})$
    end for
  end for
end for

We propose a *Wake-Sleep strategy* for training EBMs, captured in Algorithm 1, where instead of just minimizing the coupled contrastive loss function, we minimize a dynamic loss function having two decoupled phases, defined as follows:

- ***Wake Phase***: Here the loss function is the square of only the positive energy defined as $L_{wc} = E(w, x_i, y_i)^2$. In the wake phase, we are essentially minimizing the positive energy. This essentially leads to pushing down the energy of the desired answer without pulling up the energy of incorrect solutions. We view this phase as an active learning phase, where the attention of the model is solely on the correct classes and no information on incorrect answers is presented.
- ***Sleep Phase***: Here the loss function (L_{sc}) to minimize includes both the energy of the desired solution as well as the incorrect solutions. The goal is to push down the energy of correct answers and pull up the energy of all the other answers that are incorrect. This is the typical phase of training EBMs. We view this phase as a passive phase or consolidation/aggregation of the information. It is done in humans during sleep [6, 57].

5. Theoretical Analysis

In [28] it is argued that minimizing only the positive energy loss function can lead to the collapsed solution since there is no mechanism to increase the energy of incorrect solutions. The model parameters during the wake phase capture the positive class(ground truth labels) by freely minimizing their energies without any constraints. This gives the model conditioning over the correct class labels. But simply running the wake phase for the entire training procedure will deprive the model of learning the incorrect class labels, which leads to a collapsed mode. This is prevented by a longer sleep phase, which provides a margin for the model to push up the energy of the incorrect solutions for the already conditioned model obtained from the wake phase.

This procedure is explained in Algorithm 1. In order to avoid mode collapse[14, 28], where the energy manifold is a flat surface, the energy functions and the loss functions must satisfy the following conditions [28]:

Condition 1 (Necessary Condition on Energy Functions)**.** *For any sample (x_i, y_i) and model parameters w, the energy of the correct answer for x_i must be lower than the energy of the most offending incorrect answer $\bar{y_i}$ by a positive margin m:*

$$E(w, y_i, x_i) < E(w, \bar{y}_i, x_i) - m, \quad (5)$$

where the most offending incorrect answer $\bar{y_i}$ can be defined as:

$$\bar{y_i} = \textit{argmin}_{y\in\mathcal{Y}\ and\ y\neq y_i} E(w, y, x_i). \quad (6)$$

Further, in energy-based training, only the relative values of $E(w, y_i, x_i)$, denoted by E_C, and $E(w, \bar{y}_i, x_i)$, denoted by E_I, matter. Now consider a cross-section of the loss function in the 2-dimensional plane formed by these two energy values as shown in Figure 2. We can represent an arbitrary shaded region R of this slice, corresponding to all possible values of parameter w. Further, we assume the existence of at least one set of parameters w for which Condition 1 is satisfied for a single training sample (x_i, y_i). If such a w does not exist then there cannot exist any loss function whose minimization leads to Condition 1. The 2d plane can be divided into two planes P_1 and P_2 by the solid red line $E_I = E_C + m$, where m is the positive margin as stated in the Equation 5. We can now state a sufficient condition for designing the loss functions for energy-based

training, which when satisfied ensures the satisfiability of the Necessary Condition above.

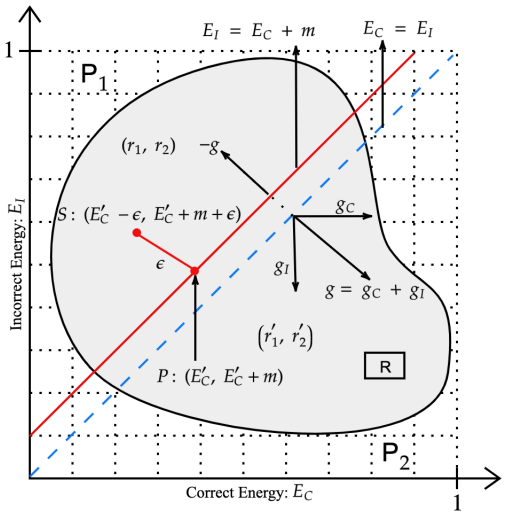


Figure 2. Direction of the negative gradient of $\mathcal{L}'$ given by vector summation of g_C and g_I in the feasible region R (grey shaded region) shows that loss decreases monotonically moving from P_2 to P_1.

Condition 2 (Sufficient Condition on Loss Function)**.** *Minimizing the loss function $\mathcal{L}$ in the feasible region R, will satisfy **necessary condition** if, there exists at least one point $(r_1, r_2) \in P_1$ such that the loss function is less than all the points (r'_1, r'_2) such that $(r'_1, r'_2) \in P_2$*

In other words, there must exist at least one point in the feasible region R intersecting the P_1, such that the value of the loss function at this point is less than the value of the loss function at all the other points in the part of *R* intersecting P_2. It can be observed that WS-EBM first conditions the model parameters on the current task by minimizing the loss in the wake cycle (L_{wc}). These are then further optimized by the loss function in the sleep cycle (L_{sc}), which consolidates knowledge on both the current task as well as the previous tasks, with less interference.

Although it is difficult to analyze the dynamic loss of WS-EBM, we present a theoretical explanation of a slightly simpler scheme, where the combined loss function can be written as a linear combination of L_{wc} and L_{sc}, given as $\mathcal{L}' = \alpha L_{wc} + \beta L_{sc}$, where α is the number of wake cycles and β is the number of sleep cycles (see Appendix for in detail discussion and proof). The actual training dynamics do not involve a straightaway linear combination that is only used here for the ease of mathematical analysis. Table 1 however clearly shows the superiority of our proposed technique over a straightforward linear combination in a practical example.

Theorem 1: *$\mathcal{L}'$ satisfies the sufficient condition for designing a loss function for EBMs.*

Proof: For $\mathcal{L}' = \alpha L_{wc} + \beta L_{sc}$, consider $L_{wc} = E(w, y_i, x_i)^2$ and $L_{sc} = E(w, x, y) + log(\sum_{\bar{y} \in \mathcal{Y}} e^{-E(w, x_i, \bar{y})})$. For any fixed parameter w and training sample (x_i, y_i), the gradient for the loss wrt to the correct energy(E_c) of the correct answer y_i and incorrect energy(E_I) of the most offending incorrect answer $\bar{y}_i$ admit the following form:

$$g_C = \frac{\partial \mathcal{L}'(w, y_i, x_i)}{\partial E_C} = 2\alpha E_C + \beta(1 - \frac{e^{-E(w, y_i, x_i)}}{\sum_{y \in \mathcal{Y}} e^{-E(w, y, x_i)}}) \tag{7}$$

$$g_I = \frac{\partial \mathcal{L}'(w, y_i, x_i)}{\partial E_I} = -\beta \frac{e^{-E(w, \bar{y}_i, x_i)}}{\sum_{y \in \mathcal{Y}} e^{-E(w, y, x_i)}} \tag{8}$$

Since α and β are the number of wake-sleep cycles respectively, they are positive. Since E_C and E_I range in $(0, 1)$, for any values of E_C, α and β, $g_C > 0$ and $g_I < 0$. The overall direction of the gradient at any point in the space of E_C and E_I is shown in Figure 2 (Figure 2 also provides an explanatory illustration for the proof). Thus we can conclude that going from P_2 to P_1, the loss decreases monotonically. Now consider a point $P = (E'_C, E'_C + m)$ lying on the margin line for which the loss is minimum. Due to monotonicity, we can conclude that

$$\mathcal{L}'(E'_C, E'_C + m) \leq \mathcal{L}'(E_C, E_I) \tag{9}$$

Now consider another point S at a distance ϵ away from the point $(E'_C, E'_C + m)$ and inside P_1, i.e., this points has coordinates $(E'_C - \epsilon, E'_C + \epsilon + m)$ and is inside P_1. From Taylor's expansion on the loss at this point S we get

$$\mathcal{L}'(E'_C - \epsilon, E'_C + m + \epsilon) = \mathcal{L}'(E'_C, E'_C + m) - \epsilon(\frac{\partial \mathcal{L}'}{\partial E_C} - \frac{\partial \mathcal{L}'}{\partial E_I}) + O(\epsilon) \tag{10}$$

From the discussion above, the second term on the right is negative, so for infinitesimally small ϵ, we have

$$\mathcal{L}'(E'_C - \epsilon, E'_C + m + \epsilon) < \mathcal{L}'(E'_C, E'_C + m) \tag{11}$$

Therefore it can be concluded that there exists at least one point in P_1 at which the loss is less than at all points in P_2. Thus $\mathcal{L}\prime$ satisfies Condition 2, which implies it satisfies Condition 1 as well. This analysis ensures that minimizing the combined loss function of WS-EBM will give us a correctly trained classifier without mode collapse.

6. Main Experiments

6.1. Experimental Setup

Out of the three settings proposed in [51] class incremental setting is the most challenging [10, 50] and we perform

LC	WS-EBM
55.32 ± 0.86	57.86 ± 0.03

Table 1. Comparison of Linear Combination(LC) with $\alpha = 1$ and $\beta = 10$ vs our proposed WS-EBM on SplitMNIST data set.

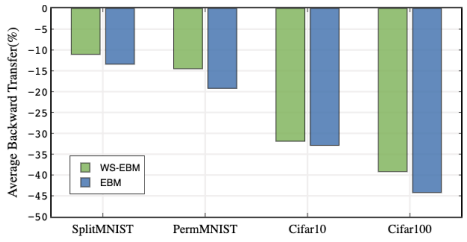
all our experiments in this scenario. Each experiment was performed 10 times with different random seeds. We report the mean $\pm$ Standard Error of Mean (SEM). The number of iterations is kept fixed at 2000 for training EBMs under traditional setting. To compare the WS-EBM, we keep the iterations per task to 200, with 2 wake cycles and 10 sleep cycles. This ensures that the total iterations per task are approximately the same. We ran all our experiments on NVIDIA V100 GPU to maintain consistency.

6.2. Datasets

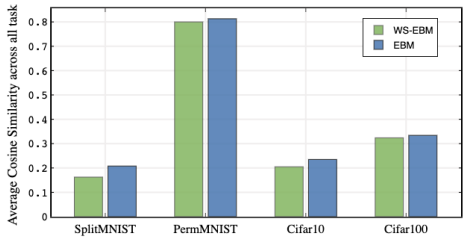
We ran our experiments on four standard continual learning benchmarks: splitMNIST [59], permutedMNIST [25], CIFAR-10 [26], and CIFAR-100 [26] data sets. The splitMNIST data set is obtained by splitting the original MNIST [27] into 5 tasks with each task having 2 classes. It has 60,000 training images and 10,000 test images. The permuted MNIST protocol has 10 tasks and each task has 10 classes. We separate CIFAR-10 into 5 tasks, each task with 2 classes. CIFAR-100 is split into 10 tasks with each task having 10 classes. This demonstrates multi-class classification. The last two data sets each have 50,000 training images and 10,000 test images.

6.3. Architecture of Energy Function

Traditional classification models only feed in x as input. In contrast, EBMs have many different ways to combine x and y in the energy function with the only requirement that $E(w, x, y) : (\mathbb{R}^D, \mathbb{N}) \rightarrow \mathbb{R}$. To compute the energy of any data x and class label y, x is sent into a small network to generate the feature $f(x)$. The label y is mapped into a same dimension feature $g(y)$ using a small learned network or a random projection. $f(x)$ and $g(y)$ are added and the output is finally sent to weight layers to generate the energy value $E(w, f(x), g(y))$.. The baseline models for computing the Energy value have been kept simple to emphasize the efficacy of our algorithm. For SplitMNIST and PermutedMNIST the baseline model architecture is similar to that in [51] and consists of a single fully connected layer with ReLU activation. For Cifar10 and Cifar100, we use the baseline model architecture as in [29], i.e., it has 5 convolutional layers connected to a fully connected layer for performing multi-class classification.



((a)) Backward Transfer



((b)) Average Orthogonality

Figure 3. Comparison of Backward Transfer [%] in Figure (a) (less negative is better) over benchmark data sets for WS-EBM vs EBM [29]. Clearly, WS-EBM has a higher BWT across all the data sets. Figure (b) shows Average Orthogonality using Cosine Similarity between gradient vectors of task t_{i+1} and t_i averaged over all tasks.

6.4. Evaluation

We evaluate the WS-EBM against available baseline models in two cases. Firstly where there is no usage of replay or any additional buffer such as standard softmax-based classifier (SBC). We report the performance of EWC [25], Online EWC (Schwarz et al., 2018), SI [59], LwF [30], MAS [1], BGD [60], and EBM [29]. For SBC, EWC, Online EWC, Online EWC, LwF on splitMNIST, permuted MNIST, and CIFAR100, we use the results reported in [51]. For BGD, we use the results from [60]. For MAS, we use the result from [41].

We also report results obtained by replay-based methods [7, 20, 32, 41, 44], which typically employ an external memory buffer. In many cases, usage of generative modeling is also often utilized [11, 47, 53]. Usually, these methods are computationally intensive, in terms of memory usage, but often give the best results since catastrophic forgetting is minimized by updating the buffer with previously seen samples. Due to the simplicity of our method, it is quite straightforward to integrate WS-EBM with the replay-based methods to further improve performance.

Methods like [22, 61] focus on learning unified classifiers by first pre-training the model for some subset of classes and then reducing forgetting. On the other hand, in our evaluation, there is no pre-training involved. Similarly methods involving transformers [12, 24] or dynamic architectures [56, 58] have been excluded from our evaluation.

Method(without replay)	SplitMNIST	PermMNIST	Cifar10	Cifar100
SBC	19.90 ± 0.02	17.26 ± 0.19	19.06 ± 0.05	8.18 ± 0.10
EWC [25]	20.01 ± 0.06	25.04 ± 0.50	18.99 ± 0.03	8.20 ± 0.09
SI [59]	19.99 ± 0.06	29.31 ± 0.62	19.14 ± 0.12	9.24 ± 0.22
LwF [30]	23.85 ± 0.44	22.64 ± 0.23	19.20 ± 0.30	10.71 ± 0.11
Online EWC [46]	19.96 ± 0.07	33.88 ± 0.49	19.07 ± 0.13	8.38 ± 0.15
BGD [60]	19.64 ± 0.03	84.78 ± 1.30	NA	NA
MAS [1]	19.50 ± 0.30	NA	20.25 ± 1.54	8.44 ± 0.27
EBM [29]	53.12 ± 0.04	87.58 ± 0.50	38.84 ± 1.08	30.28 ± 0.28
WS-EBM	$\mathbf{57.86 \pm 0.03}$	$\mathbf{88.62 \pm 0.57}$	$\mathbf{40.21 \pm 0.02}$	$\mathbf{31.71 \pm 0.24}$

Table 2. Average Accuracy [%] (higher is better) for class incremental learning over 4 data sets. All reported results are averaged over 10 runs with different random seeds.

Method(with replay)	SplitMNIST	PermMNIST	Cifar10	Cifar100
iCARL [44]	92.49 ± 0.12	91.36 ± 0.03	18.32 ± 0.21	37.83 ± 0.21
DGR [47]	90.35 ± 0.24	92.19 ± 0.09	17.21 ± 1.88	9.22 ± 0.24
GSS-Greedy [2]	84.80 ± 1.80	77.30 ± 0.50	33.56 ± 1.70	NA
A-GEM [8]	65.10 ± 3.14	83.51 ± 0.68	28.91 ± 0.02	20.38 ± 1.45
BI-R [52]	94.41 ± 0.15	NA	NA	25.81 ± 0.25
DER++ [7]	90.43 ± 1.87	83.58 ± 0.59	43.26 ± 0.76	33.91 ± 1.62
G-Dumb [41]	91.82 ± 0.51	NA	35.03 ± 0.42	24.37 ± 0.67
EBM+ER [29]	91.13 ± 0.35	94.59 ± 0.09	44.76 ± 0.73	34.07 ± 0.55
WS-EBM+ER	$\mathbf{95.81 \pm 1.36}$	$\mathbf{95.28 \pm 0.65}$	$\mathbf{45.81 \pm 0.34}$	$\mathbf{36.10 \pm 0.01}$

Table 3. Average Accuracy [%] (higher is better) for class incremental learning over 4 data sets (all methods use replay and rely on buffer size equal to 100, as recommended in [53]). All reported results are averaged over 10 runs with different random seeds.

We report the following metrics: **Average Accuracy** and **Backward Transfer** [32]. After the model finishes learning task t_i, we evaluate its test performance on all T tasks. In order to do this, we construct the matrix $R \in \mathbb{R}^{TxT}$, where $R_{i,j}$ is the test classification accuracy of the model on task t_j after observing the last sample from task t_i. Taken together, these two metrics allow us to assess how well a continual learner solves a classification problem while overcoming forgetting.

Average Accuracy: This score shows the model accuracy after training over T consecutive tasks and can be defined as: $ACC = \frac{1}{T}\sum_{i=1}^{T} R_{T,i}$.

Backward Transfer: This is the influence that learning a current task has on the performance on a previous task and is defined as: $BWT = \frac{1}{T-1}\sum_{i=1}^{T-1}(R_{T,i} - R_{i,i})$.

6.5. Results

Table 2 shows the performance results in terms of average accuracy. It can be seen that training EBMs with the wake-sleep cycles improve the average testing accuracy across all the datasets. In Table 3 we show the results of combining the WS-EBM approach with experience replay. We call it WS-EBM+ER. It can be observed that WS-EBM outperforms other baselines such as iCARL and BI-R. Figure 3 also shows the improvement in backward transfer over the existing energy-based continual learning scheme, EBM, across all data sets. Achieving zero-forgetting is very difficult for these data sets because all the tasks share at least one output layer and there is no task identifier during testing. Clearly, the strong performance on BWT indicates the efficacy of the proposed technique.

The proposed two-phase decoupled technique is the most effective in mitigating catastrophic forgetting and gives around 4% improvement in the average accuracy over prior results reported in the literature for *SplitMNIST*, 1% for **PermMNIST**, and 2% for **Cifar10** while showing a comparable performance on **Cifar100** dataset. We hypothesize that the improvement in performance in the class incremental setting is due to better knowledge consolidation through decoupled training of energy function without mode collapse. Short wake cycles decoupled with longer sleep cycles, the latter corresponding to the single-phase training regime, provide prior conditioning of the EBM to output low energy values for the correct class and decrease the interference of newer introduced classes during continual learning of new tasks. This allows the model to enlarge the margin between correct

and incorrect classes while avoiding catastrophic forgetting, as seen in the performance improvements in Table 2 and Figure 3.

Also, the wake phase provides soft conditioning on the matched or ground truth labels. To further quantize this effect, we compare the orthogonality of the gradients of the model parameters across different tasks (similar studies were done in [15, 36]). We compute gradients at $task_i$ and $task_{i-1}$ $\forall i \in T$, where T represents the total number of tasks. Cosine Similarity is utilized to compute orthogonality between gradient vectors, which are then averaged over all the tasks. A lower mean score of cosine similarity demonstrates higher orthogonality and vice-versa. Figure 3 demonstrates the average cosine similarity computed over all the tasks across all the data sets. We find that WS-EBM demonstrates higher average orthogonality across all the data sets compared to the baselines. In the Appendix, we show that WS-EBM has a lower entropy [5] as compared with EBM which signifies that our technique correlates well with generalization. Further in the appendix, we visualize the energy landscape of our proposed technique on a toy classification problem in class incremental setting and find that there are less perturbations in the evolution of energy surface in WS-EBM which shows less interference with prior learned tasks. Finally, WS-EBM was also applied with different margin loss functions [28], and it gave a superior performance as compared to the traditional training of EBM in class incremental scenarios.

7. Conclusion

In this paper, we propose a new training scheme for EBMs that rely on decoupled wake-sleep cycles. The new approach, WS-EBM, alternately switches between the phase of conditioning the model over the correct labels (short wake phase) and the phase of knowledge consolidation (long seep phase). We apply our method in the challenging class-incremental learning scenario. On multiple benchmarks, we demonstrate the superior performance of WS-EBM over a plethora of continual learning techniques, including the regular EBM training that minimizes a single loss function. Applying the concept of wake-sleep cycles can be easily extended to other domains such as regression [19] generation [13], and reinforcement learning [39], and we leave it to future works.

References

[1] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11263, 2019.

[2] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

[3] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

[4] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. Initial classifier weights replay for memoryless class incremental learning. *arXiv preprint arXiv:2008.13710*, 2020.

[5] Devansh Bisla, Jing Wang, and Anna Choromanska. Low-pass filtering sgd for recovering flat optima in the deep learning optimization landscape. In *International Conference on Artificial Intelligence and Statistics*, pages 8299–8339. PMLR, 2022.

[6] Jan Born and Ines Wilhelm. System consolidation of memory during sleep. *Psychological research*, 76:192–203, 2012.

[7] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

[8] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *ICLR*, 2019.

[9] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

[10] He Chen, Wang Ruiping, Shan S, and Chen Xilin. Exemplar-supported generative reproduction for class incremental learning. In *British Machine Vision Conference*, 2018.

[11] Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. *Advances in Neural Information Processing Systems*, 33:16481–16494, 2020.

[12] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9285–9295, 2022.

[13] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems*, 2019.

[15] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3762–3773. PMLR, 2020.

[16] Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.

[17] Robert French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3:128–135, 1999.

[18] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020.

[19] Fredrik K Gustafsson, Martin Danelljan, Goutam Bhat, and Thomas B Schön. Energy-based models for deep probabilistic regression. In *European Conference on Computer Vision*, pages 325–343. Springer, 2020.

[20] Tyler L Hayes and Christopher Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 220–221, 2020.

[21] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8): 1771–1800, 2002.

[22] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] Khurram Javed and Martha White. Meta-learning representations for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, and Bing Liu. Learnability and algorithm for continual learning. *arXiv preprint arXiv:2306.12646*, 2023.

[25] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526, 2017.

[26] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 2012.

[27] Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005.

[28] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[29] Shuang Li, Yilun Du, Gido van de Ven, and Igor Mordatch. Energy-based models for continual learning. In *Conference on Lifelong Learning Agents*, pages 1–22. PMLR, 2022.

[30] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[31] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Rmm: Reinforced memory management for class-incremental learning. *Advances in Neural Information Processing Systems*, 34:3478–3490, 2021.

[32] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

[33] Davide Maltoni and Vincenzo Lomonaco. Continuous learning in single-incremental-task scenarios. *Neural Networks*, 116:56–73, 2019.

[34] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. 24:109–165, 1989.

[35] M Jehanzeb Mirza, Marc Masana, Horst Possegger, and Horst Bischof. An efficient domain-incremental learning approach to drive in all weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2022.

[36] Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. In *International Conference on Machine Learning*, pages 15699–15717. PMLR, 2022.

[37] Sudhanshu Mittal, Silvio Galesso, and Thomas Brox. Essentials for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3513–3522, 2021.

[38] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. Distributional reinforcement learning for energy-based sequential models. *arXiv preprint arXiv:1912.08517*, 2019.

[40] Gabriel Pereyra, G. Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. *ArXiv*, abs/1701.06548, 2017.

[41] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, pages 524–540. Springer, 2020.

[42] Jathushan Rajasegaran, Munawar Hayat, Salman H Khan, Fahad Shahbaz Khan, and Ling Shao. Random path selection for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[43] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97 2:285–308, 1990.

[44] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[45] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[46] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pages 4528–4537. PMLR, 2018.

[47] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

[48] James Smith, Jonathan Balloch, Yen-Chang Hsu, and Zsolt Kira. Memory-efficient semi-supervised continual learning: The world is its own replay buffer. pages 1–8. IEEE, 2021.

[49] Jun Tani. *Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena*. Oxford University Press, 2016.

[50] Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. Topology-preserving class-incremental learning. In *Computer Vision – ECCV 2020*, pages 254–270, Cham, 2020. Springer International Publishing.

[51] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

[52] Gido M Van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):4069, 2020.

[53] Gido M. van de Ven, Zhe Li, and Andreas S. Tolias. Class-incremental learning with generative classifiers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3606–3615, 2021.

[54] Johannes von Oswald, Christian Henning, Benjamin F. Grewe, and João Sacramento. Continual learning with hypernetworks. In *International Conference on Learning Representations*, 2020.

[55] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European Conference on Computer Vision*, pages 398–414. Springer, 2022.

[56] Fu-Yun Wang, Da-Wei Zhou, Liu Liu, Han-Jia Ye, Yatao Bian, De-Chuan Zhan, and Peilin Zhao. BEEF: Bi-compatible class-incremental learning via energy-based expansion and fusion. In *The Eleventh International Conference on Learning Representations*, 2023.

[57] Yina Wei, Giri P Krishnan, Lisa Marshall, Thomas Martinetz, and Maxim Bazhenov. Stimulation augments spike sequence replay and memory consolidation during slow-wave sleep. *Journal of Neuroscience*, 40(4):811–824, 2020.

[58] S. Yan, J. Xie, and X. He. Der: Dynamically expandable representation for class incremental learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3013–3022, 2021.

[59] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017.

[60] Chen Zeno, Itay Golan, Elad Hoffer, and Daniel Soudry. Task agnostic continual learning using online variational bayes. *arXiv preprint arXiv:1803.10123*, 2018.

[61] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5867–5876, 2021.